數學傳播　48卷1期, pp. 9-18

介紹 Google PageRank 並運用於臺北捷運人流分析

洪紹鈞 · 曹怡婷 · 謝世峰

摘要: Google PageRank 是 Google 公司所使用的對其搜尋引擎結果中的網頁進行排名的一種演算法。PageRank (簡稱 PR) 的演算法是一種基於馬可夫鏈所設計而來的方法。在這篇報告中, 我們以矩陣的角度介紹 PR 演算法的原理, 用以進行台北捷運系統的人流分析。我們主要研究臺北捷運各站之間的人流情況, 並使用 PR 演算法來找出在特定時段裡最受歡迎的幾個捷運站, 以及分析其熱門的原因。

1. 簡介

Google PageRank (簡稱 PR) 是 Google 公司所使用的對其搜尋引擎搜尋結果中的網頁進行排名的一種演算法, 這個命名不但是由於對網頁 (page) 的一種排名方式, 同時演算法也是以 Google 公司創始人之一的 Larry Page 的名字來命名諧音。PR 演算法是一種基於馬可夫鏈所設計而來的方法。在這篇報告中, 我們將介紹 PR 的原理, 並且將這個方法用在台北捷運系統的人流分析。

馬可夫鏈是線性代數課程中很重要的一個應用章節, 我們簡略介紹其主要機制。現在假設我們有 n 個城市, 在此, 我們可以將這 n 個城市註記成集合 $\{1,2,\ldots,n\}$。我們也假設, 每一年這 n 個城市的居民遷徙至多一次, 城市 i 的居民遷徙至城市 j 的總人口比例 (或稱之為機率) 為 p_{ij}, 這裡的 p_{ij} 必須要是一個介於 0 與 1 之間的實數, 在此, 我們不難看出來 $p_{i1}+\cdots+p_{in}=1$。而這樣的過程就稱之為馬可夫鏈, 其中遷徙的比例並不會隨著時間而改變。假如我們將這 n^2 個比例組合成一個矩陣 $P\in M_{n\times n}$, 這個矩陣 P^T 就稱之為一個隨機矩陣 (參閱圖1)。

$$P^T=\begin{bmatrix} 0 & 0.4 & 0.6 \\ 0.5 & 0.3 & 0.2 \\ 0.5 & 0.3 & 0.2 \end{bmatrix}$$

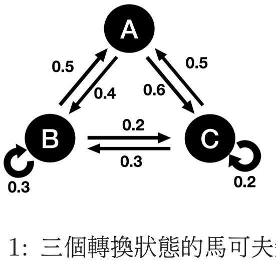

圖 1: 三個轉換狀態的馬可夫鏈

另外, 如果任一個城市 i, 他的居民所佔的總人口的比例為 u_i, 這裡的 u_i 自然是介於 0 與 1 之間的實數, 而且 $u_1 + \cdots u_n = 1$, 這樣一來, 我們也將這些 u_i 組合成一個行向量 $u = [u_1, \ldots, u_n]^T$, 這樣的行向量 $\mathbf{u}$ 就稱之為一個機率向量。我們也不難看出, $u_i p_{ij}$ 則表示前一年在城市 i 但是下一年遷徙至城市 j 的總人口比例, 所以 $u_1 p_{1j} + \cdots + u_n p_{nj}$ 則象徵過一年之後城市 j 所佔的總人口比例。類推下去, $(P^T)^k \mathbf{u}$ 表示經歷 k 年的時候, 這 n 座城市的人口分佈。當然我們會很好奇這樣長期的遷徙下去之後會發生哪些事情, 而在解釋馬可夫鏈在長時間遷徙下的表現, 我們先介紹以下定理。

定理 1.1 ([1]中的定理5.4)**.** 假設所有的$p_{ij} > 0$, 則以下兩件事情成立:

(a) 存在唯一的機率向量$\mathbf{v} \in \mathbb{R}^n$使得 $P^T \mathbf{v} = \mathbf{v}$;

(b) 對於任意的機率向量$\mathbf{u} \in \mathbb{R}^n$, 以下的極限皆成立 $\lim_{k\to\infty}(P^T)^k \mathbf{u} = \mathbf{v}$。

從定理 1.1 的結論 (b) 我們可以看出來, 無論一開始人口如何分佈, 長久時間下來, 一定會趨近於平衡狀態 $\mathbf{v}$。我們可以這樣解讀定理 1.1 的結論 (b), 假設 $v_1 = 0.4$ 以及 $v_2 = 0.1$, 這就表示這個區域人民長期遷徙下來, 最終在城市1的人口佔了總人口的 40%, 而城市 2 的人口佔了總人口的 10%, 相對來說, 有更多的人想要居住在城市 1 而不是城市 2, 定理 1.1 所說明的就是這種城鄉差異。

在 2000 年前後, Google 公司以「投票機制」與「阻尼係數」兩種概念, 巧妙地將網際網路中各個網頁的連結關係建構出一個超大型的隨機矩陣 P^T, 在 2023 年全球網際網路中的網頁總數大約是 $n \approx 10^{10}$ (取自 [3])。網際網路中的每一個網頁也都可以被 1 至 n 中的一個正整數代表。由於網頁與網頁之間會有 *超連結*的溝通, 因而 p_{ij} 則用來表示一個使用者停留在網頁 i 的時候連往網頁 j 的機率。Google 公司便在背後計算定理 1.1 中的向量 $\mathbf{v}$, 我們已經知道 v_j 可以表示最終所有網路使用者停留在網頁 j 的機率。因而, Google 公司便將 v_j 稱之為網頁 j 的 PR, 在此我們不難想像, v_j 越高經常表示這個網頁越受關注。

在這一篇報告中, 我們也仿照 Google 計算 PR 的方式, 利用公開資訊將台北捷運製作成一個 119×119 的隨機矩陣, 藉由不同時段的人流資訊分析定義 p_{ij} 表示一名乘客由捷運站 i 至捷運站 j 的機率, 這樣一來, 我們也可以計算該時段各站的 PR 值。

在以下的第二節, 我們首先介紹定理 1.1 的原理, 藉此, 我們也將介紹 Google 如何將這些網頁之間的關係製作成一個超大型的隨機矩陣 P, 有了 P 之後, 我們也將介紹隨機向量 $\mathbf{v}$ 的計算方法。這樣一來, 我們就可以介紹 Google 如何將網頁的關係以一個隨機矩陣 P 表示, 並由我們上述的計算方法算出每個網頁的 PR 值。在本文第三節, 我們仿照 Google 計算 PR 值的方法, 計算 2021 幾個時段的台北捷運的 PR 值。我們在第四節總結這份報告。

2. 馬可夫鏈與Google PageRank

2.1. 馬可夫鏈的機制與定理 1.1 的原理

我們先來介紹定理 1.1 的原理。在此, 我們僅考慮矩陣 P 爲一個 2×2 矩陣的情形。在這個設定之下, P^T 可以寫成

$$P^T=\begin{bmatrix}\alpha & \beta\\ 1-\alpha & 1-\beta\end{bmatrix}; \tag{2.1}$$

這裡的 α 與 $\beta\in(0,1)$。而對於更一般的情況的證明, 由於牽涉到更多的矩陣分析的知識, 我們在此略過, 讀者可以參考書本 [4] 中 Perron-Frobenius Theorem 的證明。爲了方便說明原理, 我們先定義以下兩個名詞。

定義 2.1. 假設 A 是一個 $n\times n$ 的實數矩陣。如果有一實數 λ 以及一非零向量 $\mathbf{u}\in\mathbb{R}^2$ 滿足以下性質

$$A\mathbf{u}=\lambda\mathbf{u}.$$

這時候, λ 則稱爲矩陣 A 的一個特徵值, $\mathbf{u}$ 則稱之爲對應於 λ 的 特徵向量。

我們反觀定理 1.1(a), 其實這裡是在說明矩陣 P^T 有特徵值 1, 而且他對應的特徵向量其實是一個隨機向量。現在, 我們先來證明這件事。

命題 2.1. 方程式 (2.1) 中的矩陣 P^T 的特徵值與特徵向量有以下性質:

(a) 兩個特徵值分別爲 $\lambda_1=1$ 與 $\lambda_2=\alpha-\beta$;

(b) $\mathbf{v}=\begin{bmatrix}\beta\\ 1-\alpha\end{bmatrix}/(1-\alpha+\beta)$ 爲一對應 $\lambda_1=1$ 的特徵向量, 這裡的向量 $\mathbf{v}$ 是諸多特徵向量中唯一的一個機率向量;

(c) $\mathbf{w}=\begin{bmatrix}1\\ -1\end{bmatrix}$ 是對應 $\lambda_2=\alpha-\beta$ 的特徵向量。

證明: 假設 λ 爲矩陣 P^T 的特徵值, 由特徵值的定義(見定義 2.1), 我們可以得知 $\begin{vmatrix}\alpha-\lambda & \beta\\ 1-\alpha & 1-\beta-\lambda\end{vmatrix}=0$。化簡後, 我們可以得知上述方程式的根即爲 $\lambda_1=1$ 與 $\lambda_2=\alpha-\beta$, 這便是這個命題結論 (a) 的證明了。另一方面, 再經由定義 2.1 中特徵向量的定義, 我們可以知道在特徵值 λ 已經知道的前提之下, 對應 λ 的特徵向量會滿足以下方程式

$$\begin{bmatrix}\alpha-\lambda & \beta\\ 1-\alpha & 1-\beta-\lambda\end{bmatrix}\begin{bmatrix}u_1\\ u_2\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}. \tag{2.2}$$

分別將 $\lambda_1 = 1$ 與 $\lambda_2 = \alpha - \beta$ 代入方程式 (2.2) 之後, 所有對應 λ_1 的特徵向量都落於以下空間 $E_{\lambda_1} = \{c\mathbf{v}|\ c \in \mathbb{R}\}$; 所有對應 λ_2 的特徵向量都落於以下空間 $E_{\lambda_2} = \{c\mathbf{w}|\ c \in \mathbb{R}\}$。這樣我們就證明了命題 (b) 與 (c) 之中特徵向量的形式。關於上述的兩個空間 E_{λ_1} 與 E_{λ_2}, 有時候我們也會把他們稱之爲 特徵空間。另外, 假設 $\mathbf{u}$ 是 E_{λ_1} 中的向量, 那麼 $\mathbf{u} = c\mathbf{v}$。因此, $u_1 + u_2 = c$。這表示 $\mathbf{v}$ 是 E_{λ_1} 中唯一的機率向量。 □

在命題 2.1 中, 我們推導出兩個 P^T 的特徵向量 $\mathbf{v}$ 與 $\mathbf{w}$, 在以下命題, 我們會看到, 任何一個機率向量 $\mathbf{u}$ 都可以是這兩個向量的線性組合, 此外, 在 $\mathbf{v}$ 方向的組合係數一定是 1。

命題 2.2. 假設 $\mathbf{u} \in \mathbb{R}^2$ 爲一機率向量。則我們可以用命題 2.1 中的特徵向量 $\mathbf{v}$ 與 $\mathbf{w}$ 將 $\mathbf{u}$ 寫成以下的線性組合

$$\mathbf{u} = \mathbf{v} + \frac{(u_1 - u_2) + (1-\alpha-\beta)/(1-\alpha+\beta)}{2}\mathbf{w}。 \tag{2.3}$$

由於命題 2.2 的推導只是單純的代入驗算過程, 因此我們在此省略。現在我們來說明第三個性質, 也就是長期遷徙過後的人口的分佈性質。

命題 2.3. 假設 $\mathbf{u} \in \mathbb{R}^2$ 爲一機率向量。則 $\lim\limits_{k\to\infty}(P^T)^k\mathbf{u} = \mathbf{v}$。

證明: 由命題 2.2, 我們可以知道 $\mathbf{u}$ 是 $\mathbf{v}$ 與 $\mathbf{w}$ 的線性組合。在此我們留意一件事, 由定義 2.1 中特徵向量的定義, 我們可以得出以下兩個等式:

$$(P^T)^k\mathbf{v} = \mathbf{v},\ (P^T)^k\mathbf{w} = \lambda_2^k\mathbf{w}。 \tag{2.4}$$

此外, 由於 $\alpha, \beta \in (0,1)$, 我們也知道 $|\lambda_2| = |\alpha - \beta| < 1$。現在我們將方程式 (2.3) 改寫成以下形式 $\mathbf{u} = \mathbf{v} + c\mathbf{w}$。接著, 我們在此方程式的等號兩側的左邊都乘以矩陣 $(P^T)^k$, 並且使用等式 (2.4) 與代換, 我們便可以得到一個新的方程式

$$(P^T)^k\mathbf{u} = \mathbf{v} + c\lambda_2^k\mathbf{w}。 \tag{2.5}$$

因爲 $|\lambda_2| < 1$, 故 $\lim_{k\to\infty} \lambda_2^k = 0$, 所以方程式 (2.5) 便可以得出本命題的結論。 □

定理 1.1 **的證明**: 當 $n = 2$ 的時候, 定理 1.1 即爲命題 2.1 (b) 的推論。而命題 2.3 則直接說明了定理 1.1(c)。 □

2.2. Google PageRank

全球有大約 $n \approx 10^{10}$ 個網頁, 而這些網頁與網頁之間存在著超連結溝通, 現在我們將這

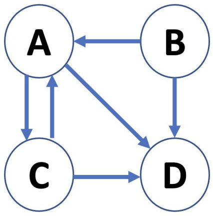

圖 2: 四個網頁關係圖

些網頁以正整數 1 至 n 排序。現在我們來定義鄰接矩陣$E \in M_{n\times n}$:

$$e_{ij} = \begin{cases} 1 & \text{如果頁面 } i \text{ 中有一個超連結連往頁面 } j; \\ 0 & \text{如果頁面 } i \text{ 中沒有超連結連往頁面 } j。\end{cases}$$

以圖 2 爲例, E^T 可以表示成 $E^T = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$。現在我們令

$$s_i = \text{頁面}i\text{所連結出去的總頁面數} = \text{矩陣}E^T\text{第}i\text{列的總和}, \tag{2.6}$$

以及令阻尼係數(damping factor)

$$\gamma = 0.85。$$

基於上述的假設, Google 對於一個網路使用者的行爲有以下兩個假設:

假設一: 當一個網頁使用者停留在一個有對外連結的網頁 i, 即 $s_i > 0$ 的情況。這位使用者有 γ 的可能性從網頁 i 隨機點選連結, 以及有 $1-\gamma$ 的可能性從整個網際網路隨機開啟一個新的網頁。所以, 這個時候連往另一個網頁j的機率則爲

$$p_{ij} = \gamma\frac{e_{ij}}{s_i} + (1-\gamma)\frac{1}{n}。 \tag{2.7}$$

假設二: 當一個網頁使用者停留在一個沒有對外連結的網頁 i, 即 $s_i = 0$ 的情況。 這位使用者會從整個網際網路隨機開啟一個新的網頁。所以, 這個時候連往另一個網頁 j 的機率則爲 $p_{ij} = \frac{1}{n}$。

我們在此簡略以「投票機制」與「阻尼係數」兩種觀點說明上述的假設一。在方程式 (2.7) 中, 我們將 p_{ij} 分成兩項: $\gamma\frac{e_{ij}}{s_i}$ 與 $(1-\gamma)\frac{1}{n}$。$\frac{e_{ij}}{s_i}$ 即所謂的投票機制。現在我們想像每次在網頁上的點擊就是一個投票行爲, 網頁 i 一共有 s_i 個連結通往其他網頁, 但是我們只會點擊一下,

因此平均起來這每個網頁會平分出 $\frac{1}{s_i}$ 張票。而 $\frac{e_{ij}}{s_i}$ 則是因爲取決於網頁 j 是不是一個網頁 i 所連出來的。而這裡的阻尼係數 γ 所扮演的角色在於, 一個網路使用者通常不一定會安分地從網頁中點選一個連結連出去, 經常也會從這些連結之外再重新開啟一個新的網頁, 而這兩類行爲的機率分別是 γ 與 $(1-\gamma)$。而阻尼係數的機制讓矩陣 P^T 可以變成一個全滿的矩陣, 因爲任何兩個網頁都有相連的可能。如此一來, 定理 1.1 條件就滿足了, 我們隨後也可以使用定理 1.1 結論。以圖 2 爲例, $P^T=\begin{bmatrix}\frac{3}{80}&\frac{37}{80}&\frac{37}{80}&\frac{1}{4}\\ \frac{3}{80}&\frac{3}{80}&\frac{3}{80}&\frac{1}{4}\\ \frac{37}{80}&\frac{3}{80}&\frac{3}{80}&\frac{1}{4}\\ \frac{37}{80}&\frac{37}{80}&\frac{37}{80}&\frac{1}{4}\end{bmatrix}$。截至目前爲止, 我們先整理以下結論。

定理 2.4. 假設一與假設二所定義出來的矩陣P^T爲一全滿的$n\times n$隨機矩陣。

這個定理的驗證很單純, 因此我們略過。文章 [6] 裡面也有對於隨機矩陣的相關知識。而定理 2.4 告訴我們的事情是在假設一二下 Google 所架構的矩陣 P^T 是一個隨機矩陣, 同時, 這裡的 p_{ij} 表示一位網路使用者從網頁 i 連結至網頁 j 的機率, 這點, 其實還告訴我們另一件事, 即 P^T 會滿足第一節所介紹的定理 1.1 的兩個性質。基於這件事, 我們可以來定義所謂的 PageRank。

定義 2.2. 令 P 表示假設一與假設二所定義出來的隨機矩陣, 而向量 $\mathbf{v}$ 表示定理 1.1 中的機率向量。這時候, v_i 則稱之爲網頁 i 的 PageRank。

在此, 我們不妨先來總結先前幾個章節所完成的事情。我們引用定理 1.1, 將平衡狀態向量 $\mathbf{v}$ 的第 i 個分量 i 命名爲網頁 i 的 PR, 同時, 我們也可以知道, PR 也可以表示網路使用者在這個網頁上停留的機率, 值越高代表會有更多人來閱讀這個網頁。從另一個觀點來看, 在定義 2.2 中 PageRank 的定義其實是一個已知數學存在性時下的定義, 也就是說, 我們證明了他的存在, 但是我們並不知道要如何求得這個數, 原因在於一旦我們想要解方程式 $P^T\mathbf{v}=\mathbf{v}$, 尤其當 P^T 的大小爲 $n\approx 10^{10}$ 的時候, 這個問題將非常困難。在下個小節, 我們將介紹其計算方法。

2.3. Power 方法: PageRank 的計算過程

PR 的計算源自於所謂的 Power 方法 (power method)。在計算上, 它主要是用來處理一個方陣的最大的實數特徵值與特徵向量的計算方法。我們從定理 1.1 說起。定理 1.1 (b) 的極限成立意味著, 假如我們一開始隨機地選取了一個 $\mathbb{R}^n$ 中的機率向量, 稱之爲 $\mathbf{u}$, 當 k 很大的時候, $(P^T)^k\mathbf{u}\approx\mathbf{v}$, 基於這樣的想法, 我們可以採用表 1 中的 Power 方法。

在演算法的第 5 條中, 我們做了一個校正動作, 在計算上稱之爲常規化(normalization)。另一方面, 我們也會發現, Power 方法並不像傳統的數學計算直接把解答求出來, 相對的, 它是以數列收斂的型態去計算一個真實解的近似值。

Table 1: Power方法(演算法)

Input: $Tolerance$;

Let $\mathbf{u}_0 = \begin{bmatrix} 1/n \\ \vdots \\ 1/n \end{bmatrix}, \mathbf{u}_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$;

While $\|\mathbf{u}_1 - \mathbf{u}_0\| \geq Tolerance$

    Update $\mathbf{u}_1 = P^T\mathbf{u}_0$;

    Update $\mathbf{u}_1 = \mathbf{u}_1/(u_1(1) + \cdots + u_1(n))$;

    Update $\mathbf{u}_0 = \mathbf{u}_1$;

End While;

Output $\mathbf{u}_1$.

定理 2.5. 假設 P^T 爲一 $n \times n$ 全滿的隨機矩陣, 向量數列 $\{\mathbf{u}_k\}$ 爲 Power 方法所迭代出來的每一個向量。令 λ_2 表示 P^T 的特徵值中比 1 小但是絕對值最大的特徵值, 則存在一個正數 M 使得 $\|\mathbf{u}_k - \mathbf{v}\| \leq M|\lambda_2|^k$。

定理 2.5 的一般性證明其實非常的困難, 我們在此一樣請讀者參閱書本 [2]。以下, 我們一樣以 $n = 2$ 爲例來說明其中的原理。

證明: 首先, 由命題 2.1(a), 我們得知 $\lambda_2 = \alpha - \beta$, 此外, 由於 $\mathbf{u}_k = (P^T)^k\mathbf{u}_0$, 由推導命題 2.3 時候的方程式 (2.5), 我們便可以得到 $\|\mathbf{u}_k - \mathbf{v}\| \leq |c|\|\mathbf{w}\||\lambda_2|^k$。這裡的 $|c|\|\mathbf{w}\|$ 即是定理所敘述的常數M。 □

而在定理 2.5 中, 我們也可以看到另一件事, 假設給定好一個誤差的容忍值 Tol, 因此我們需要一個迭代步數 k 讓 $M|\lambda_2|^k \leq Tol$, 那麼 Power 方法大約需要

$$k \approx \frac{\log Tol - \log M}{\log|\lambda_2|}。$$

這裡我們可以看到, 如果 $|\lambda_2|$ 越小的話, 那麼 Power 方法收斂的速度就越快。

3. 台北捷運系統人流分析的應用

這一節我們將 PR 概念應用在台北捷運系統的人流分析上。捷運系統跟網際網路連結有一個相似之處, 就是每一位乘客一定都是從某個站搭乘至另一個站, 這裡的另一個站包含原先出發的捷運站。因此, 如果想做一個這樣的分析, 首先我們就需要爲台北捷運系統建構一個隨機矩陣 P^T。要進行這個工作, 我們先來看一部份「政府資料開放平臺」所提供的「臺北捷運各站分時進出量統計」的數據 [5]。首先我們擷取 2021/12/31 零時爲例, 部分的資料呈現如表 2 所示。

Table 2: 2021/12/31 0時公館到各站人數的部分呈現

日期	時段	進站	出站	人次
2021-12-31	0	公館	七張	36
2021-12-31	0	公館	大坪林	55
2021-12-31	0	公館	景美	48
2021-12-31	0	公館	萬隆	37
2021-12-31	0	公館	公館	14
2021-12-31	0	公館	台電大樓	18
2021-12-31	0	公館	古亭	29
2021-12-31	0	公館	中正紀念堂	4
2021-12-31	0	公館	小南門	11
2021-12-31	0	公館	頂溪	19

基於這一份資料, 我們可以將 2021/12/31 當日從 A 站搭乘至 B 站的所有的人數計算出來。在紀錄資料的當下, 台北捷運系統一共有 $n = 119$ 個捷運站, 因此, 我們首先可以列出一個 119×119 的矩陣 E^T, 如下: $E^T = \begin{bmatrix} 66 & 197 & \cdots & 3 \\ 56 & 204 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 152 & 180 & \cdots & 67 \end{bmatrix}$。如同上一節所介紹的, 我們在此讓

$$e_{ij} = 2021/12/31 \text{ 全日從 } i \text{ 站搭乘至 } j \text{ 站總人數。}$$

與方程式 (2.6) 類似, 我們並且令

$$\begin{aligned} s_i &= \text{捷運站 } i \text{ 所進站的總人數} = \text{捷運站 } i \text{ 搭乘至各站的人數和} \\ &= \text{矩陣 } E^T \text{ 第 } i \text{ 列的總和,} \end{aligned}$$

由於並沒有某個 s_i 爲 0, 也就是這個時段整個台北捷運系統的所有站都有人進站, 因而, 我們基於假設一, 加進去適當的阻尼係數 $\gamma = 0.85$ 的影響之後, 令

$$p_{ij} = \gamma\frac{e_{ij}}{s_i} + (1-\gamma)\frac{1}{n} = \text{捷運站 } i \text{ 搭乘至捷運站 } j \text{ 的機率。}$$

這就組成隨機矩陣。下一步, 我們同樣使用 2.3 節所介紹的 Power 方法計算每個站的 PR 值, 在計算之前, 原本預期是 101、西門、東區等地方, 應該會是最熱鬧人流最多的。然而得到的結果卻是永寧站以及土城站, 是當天出站 PR 值最高的 (見圖 3)。所以我們發現跨年當天的永寧站及土城站, 人流出站比例之高勢必受其他原因影響。

爲了解答這個問題, 我們將跨年當天分時段拆成下午五點到晚上十二點來進行細部分析, 結果顯示出來這個時段, 市政府站及西門站等跨年景點人流出站相對較高(見圖 4)。經我們調查過後, 發現原來永寧、土城位在板南線上, 是桃園、三峽、大溪等地直達台北市中心最方便的捷運站, 因此該區會是很重要的轉乘點。

圖 3: 2021 跨年夜整天藍線捷運 PR

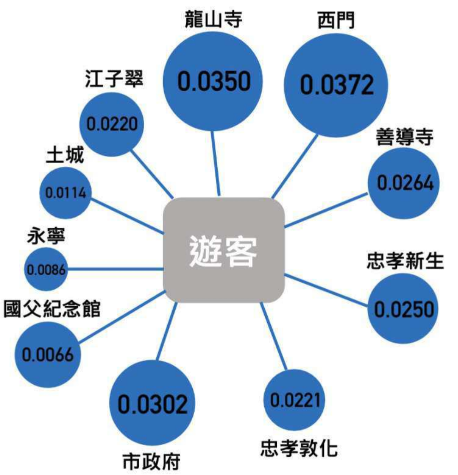

圖 4: 2021 跨年夜下午五點到十二點 PR

除此之外, 在跨年夜晚這段期間大量人潮聚集至西門、龍山寺等萬華區, 該地區爲餐廳和商店的高密集地區。市政府站是第三名, 因爲101煙火及跨年演唱會會場就是舉辦在市政府站, 前後兩站國父紀念館站和永春站也是 PR 值偏高的區域, 可見12/31晚間時段市政府該區域是捷運人潮流動的主要方向。同樣分析農曆過年的時段 (見圖 5), 我們可以看到 PR 值最高的捷運站是七張站。七張那鄰近大坪林、景美等, 在經過了解之後, 原來在該區因爲房屋價格合理, 因此有大量的上班族以及學生會選擇在這一個區域定居或租屋。即使是過年, 但也因爲很多公司初二後就開工, 所以新店一帶的人口還是會來得很高。

圖 5: 過年時段PR

圖 6: 平日時段PR

最後我們分析平日出站比例較高的捷運站, 同樣拿跟過年一樣的捷運站來做分析, 結果最

高同樣是七張 (見圖 6)。從其他天合併來看, 新店區是捷運人口流動最密集的區域, 其前後新店區公所、七張、大坪林 PR 值也同樣非常高, 整體合併來看, 新店一整區的人口流動更是非常的密集。

4. 結論

這一篇報告中, 我們仿照 Google PR 計算出台北捷運系統各站的 PR 值, 這相當於每個站的熱門程度。從上述的人流分析中, 我們可以做一個簡單的推薦系統給遊客建議, 假設一名旅客由捷運站 i 進站, 則 p_{ij} 表示預計從這個站搭車到捷運站 j 的機率, 因此, v_ip_{ij} 最大的話則表示到這個站的人很多而且這個站很熱門, 因此, 這樣我們則令

$$f(j)=\arg\max\{v_ip_{ij}|\ i\in\{1,\ldots,n\}\}$$

來表示從捷運站 j 出發的推薦值。例如: 今天遊客決定先到淡水去玩, 我們就可以推薦他到中正紀念堂, 避免接下來不知道去哪的問題。

這樣的推薦方式相對粗糙, 所以加入各種適當的考量, 例如景點、食物、娛樂等的人潮, 來給予相關性最高的推薦地方。有了這樣的推薦系統, 將能給予許多來旅遊的遊客建議, 來安排一系列的行程。

5. 致謝

作者要感謝台師大黃聰明教授與台大資工所廖家緯同學在本文撰寫的過程中提供了寶貴的意見、指導及鼓勵。

參考文獻

1. L. Friedberg, A. Spence, and S. Insel, *Elementary Linear Algebra A Matrix Approach*, Pearson, 2014.
2. G. H. Golub and C. F. van Loan, *Matrix Computations*, JHU Press, fourth edition, 2013.
3. K. Haan, K., Top Website Statistics For 2023, https://www.forbes.com/advisor/business/software/website-statistics/ .
4. R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.
5. 政府資料開放平台。臺北捷運各站分時進出量統計。https://data.gov.tw/dataset/128506 .
6. 洪萬生, 蘇惠玉, 林倉億。轉移矩陣二三事: 馬可夫鏈穩定狀態的判別。HPM 通訊, 17(7), 2-6, 2014。

—本文作者洪紹鈞, 曹怡婷就讀國立台灣師範大學; 謝世峰任教於國立台灣師範大學—